

Knowledge Discovery and Data Mining

Unit # 18

Sajjad Haider

Fall 2012

1

Mining Text Data

Data Mining / Knowledge Discovery



Structured Data

HomeLoan (
 Loanee: Frank Rizzo
 Lender: MWF
 Agency: Lake View
 Amount: \$200,000
 Term: 15 years
)

Multimedia



Free Text

Frank Rizzo bought his home from Lake View Real Estate in 1992. He paid \$200,000 under a 15-year loan from MW Financial.

Hypertext

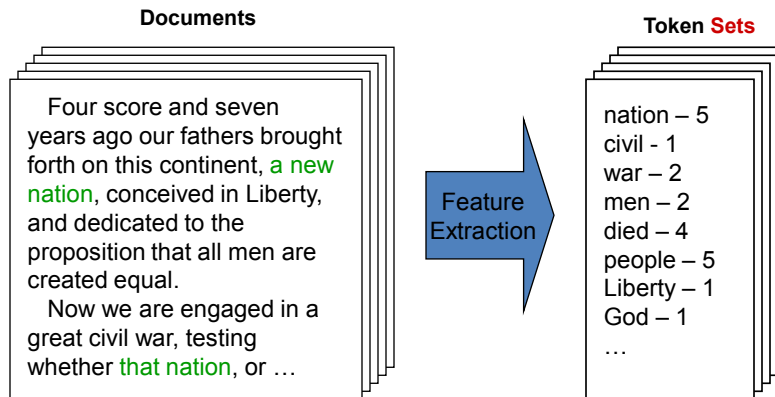
[Frank Rizzo](#) Bought [this home](#) from [Lake View Real Estate](#) In **1992**.
 <p>...

Sajjad Haider

Fall 2012

2

Bag-of-Tokens Approaches



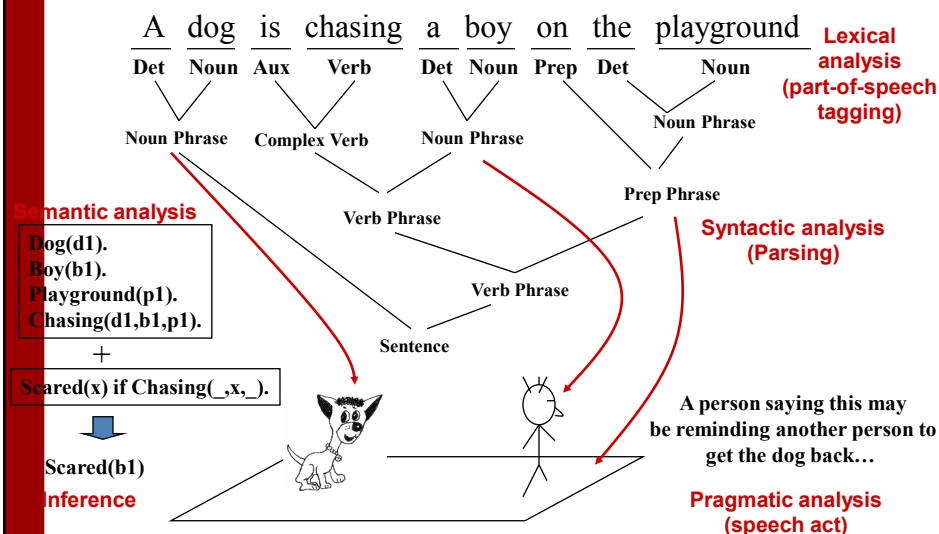
Loses all order-specific information!
Severely limits context!

Sajjad Haider

Fall 2012

3

Natural Language Processing



Sajjad Haider

Fall 2012

4

General NLP—Too Difficult!

- Word-level ambiguity
 - “**design**” can be a noun or a verb (Ambiguous POS)
 - “**root**” has multiple meanings (Ambiguous sense)
- Syntactic ambiguity
 - “**natural language processing**” (Modification)
 - “**A man saw a boy with a telescope.**” (PP Attachment)
- Anaphora resolution
 - “**John persuaded Bill to buy a TV for himself.**”
(himself = John or Bill?)
- Presupposition
 - “**He has quit smoking.**” implies that he smoked before.

**Humans rely on context to interpret (when possible).
This context may extend beyond a given document!**

Sajjad Haider

Fall 2012

5

Text Databases and IR

- Text databases (document databases)
 - Large collections of documents from various sources: news articles, research papers, books, digital libraries, e-mail messages, and Web pages, library database, etc.
 - Data stored is usually *semi-structured*
 - Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data
- Information retrieval
 - A field developed in parallel with database systems
 - Information is organized into (a large number of) documents
 - Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents

Sajjad Haider

Fall 2012

6

Information Retrieval

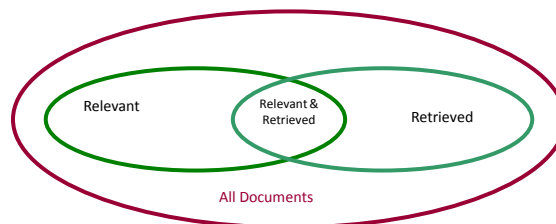
- Typical IR systems
 - Online library catalogs
 - Online document management systems
- Information retrieval vs. database systems
 - Some DB problems are not present in IR, e.g., update, transaction management, complex objects
 - Some IR problems are not addressed well in DBMS, e.g., unstructured documents, approximate search using keywords and relevance

Sajjad Haider

Fall 2012

7

Basic Measures for Text Retrieval



- **Precision:** the percentage of retrieved documents that are in fact relevant to the query (i.e., “correct” responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- **Recall:** the percentage of documents that are relevant to the query and were, in fact, retrieved

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Sajjad Haider

Fall 2012

8

Information Retrieval Techniques

- Basic Concepts
 - A document can be described by a set of representative keywords called **index terms**.
 - Different index terms have varying relevance when used to describe document contents.
 - This effect is captured through the **assignment of numerical weights to each index term** of a document. (e.g.: frequency, tf-idf)
- DBMS Analogy
 - Index Terms → **Attributes**
 - Weights → **Attribute Values**

Keyword-Based Retrieval

- A document is represented by a string, which can be identified by a set of keywords
- Queries may use **expressions** of keywords
 - E.g., car **and** repair shop, tea **or** coffee, DBMS **but not** Oracle
 - Queries and retrieval should consider **synonyms**, e.g., repair and maintenance
- Major difficulties of the model
 - **Synonymy**: A keyword *T* does not appear anywhere in the document, even though the document is closely related to *T*, e.g., data mining
 - **Polysemy**: The same keyword may mean different things in different contexts, e.g., mining

Similarity-Based Retrieval in Text Data

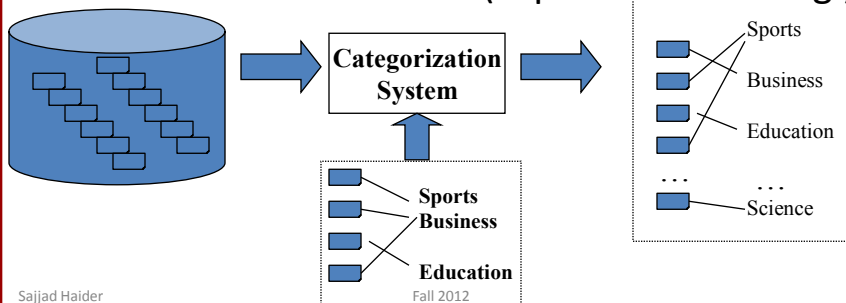
- Finds similar documents based on a set of common keywords
- Answer should be based on the degree of relevance based on the nearness of the keywords, relative frequency of the keywords, etc.
- Basic techniques
- Stop list
 - Set of words that are deemed “irrelevant”, even though they may appear frequently
 - E.g., *a, the, of, for, to, with*, etc.
 - Stop lists may vary when document set varies

Document Clustering

- Motivation
 - Automatically group related documents based on their contents
 - No predetermined training sets or taxonomies
 - Generate a taxonomy at runtime
- Clustering Process
 - Data preprocessing: remove stop words, stem, feature extraction, lexical analysis, etc.
 - Hierarchical clustering: compute similarities applying clustering algorithms.
 - Model-Based clustering (Neural Network Approach): clusters are represented by “exemplars”. (e.g.: SOM)

Text Categorization

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard classification (supervised learning.)



Sajjad Haider

Fall 2012

13

Applications

- News article classification
- Automatic email filtering
- Webpage classification
- Word sense disambiguation
-

Sajjad Haider

Fall 2012

14

Categorization Methods

- Manual: Typically rule-based
 - Does not scale up (labor-intensive, rule inconsistency)
 - May be appropriate for special data on a particular domain
- Automatic: Typically exploiting machine learning techniques
 - K-nearest neighbor (KNN)
 - Decision-tree (learn rules)
 - Neural Networks (learn non-linear classifier)
 - Support Vector Machines (SVM)
 - Naïve Bayes classifier

Society

Nodes: individuals

Links: social relationship
(family/work/friendship/etc
.)



S. Milgram (1967)

John Guare

Six Degrees of Separation

Social networks: Many individuals with diverse
social interactions between them.

Communication networks

The Earth is developing an electronic nervous system, a network with diverse nodes and links are

-computers

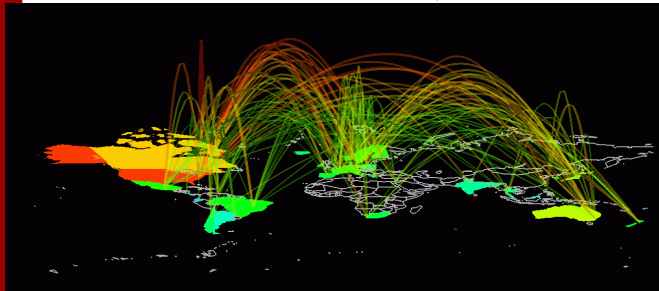
-routers

-satellites

-phone lines

-TV cables

-EM waves



Communication networks: Many non-identical components with diverse connections between them.

17

Data Mining Applications

- Data mining is an interdisciplinary field with wide and diverse applications
 - There exist nontrivial gaps between data mining principles and domain-specific applications
- Some application domains
 - Financial data analysis
 - Retail industry
 - Telecommunication industry
 - Biological data analysis

Data Mining for Financial Data Analysis

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
 - View the debt and revenue changes by month, by region, by sector, and by other factors
 - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
 - feature selection and attribute relevance ranking
 - Loan payment performance
 - Consumer credit rating

Sajjad Haider

Fall 2012

19

Financial Data Mining

- Classification and clustering of customers for targeted marketing
 - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
 - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
 - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Sajjad Haider

Fall 2012

20

Data Mining for Retail Industry

- Retail industry: huge amounts of data on sales, customer shopping history, etc.
- Applications of retail data mining
 - Identify customer buying behaviors
 - Discover customer shopping patterns and trends
 - Improve the quality of customer service
 - Achieve better customer retention and satisfaction
 - Enhance goods consumption ratios
 - Design more effective goods transportation and distribution policies

Sajjad Haider

Fall 2012

21

Data Mining for Telecomm. Industry

- A rapidly expanding and highly competitive industry and a great demand for data mining
 - Understand the business involved
 - Identify telecommunication patterns
 - Catch fraudulent activities
 - Make better use of resources
 - Improve the quality of service
- Multidimensional analysis of telecommunication data
 - Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

Sajjad Haider

Fall 2012

22

Data Mining for Telecomm. Industry (2)

- Fraudulent pattern analysis and the identification of unusual patterns
 - Identify potentially fraudulent users and their atypical usage patterns
 - Detect attempts to gain fraudulent entry to customer accounts
 - Discover unusual patterns which may need special attention
- Multidimensional association and sequential pattern analysis
 - Find usage patterns for a set of communication services by customer group, by month, etc.
 - Promote the sales of specific services
 - Improve the availability of particular services in a region
- Use of visualization tools in telecommunication data analysis

DM Courses Available Online

- MIT
 - <http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/>
- Cleveland State University Ohio
 - <http://academic.csuohio.edu/fuy/EEC%20525/syllabus.html>
- Virginia Tech
 - <http://people.cs.vt.edu/~ramakris/Courses/CS6604/lectures.html>

DM Courses Available Online (Con'td)

- Worcester Polytechnic Institute
 - http://web.cs.wpi.edu/~cs525d/f09/schedule_cs525_f09.html
- Central Connecticut State University
 - http://www.cs.ccsu.edu/~markov/ccsu_courses/580Syllabus.html
- Temple University
 - http://www.cs.ccsu.edu/~markov/ccsu_courses/580Syllabus.html

Course Recap

Course Objectives

- Know the knowledge discovery process
- Understand the different categories of algorithms
- Be able to judge which algorithms fit different problems
- Have practical experience choosing algorithms for a specific problem
- Have practical experience working in technical teams
- Have practical experience executing data mining projects
- Have practical experience using open source data mining software

Sajjad Haider

Fall 2012

27

Course Outline

- Data Preparation (Cleansing, Normalization, Transformation)
- Classification
 - Application of different algorithms/techniques
 - Feature Selection/Dimension Reduction
 - Model Evaluation
- Clustering
 - Various Algorithms
 - Model Evaluation (External vs. Internal metrics)
- Association Rules

Sajjad Haider

Fall 2012

28